OCEAN SCIENCES DATA RESCUE INVENTORY

Anthony W. Isenor and Fasil Woldegeorgis

Ocean Sciences Division
Maritimes Region
Fisheries and Oceans Canada

Bedford Institute of Oceanography
P.O. Box 1006
Dartmouth, Nova Scotia
Canada B2Y 4A2

2000

Canadian Technical Report of Hydrography and Ocean Sciences 205

Fisheries and Oceans / Pêches et Océans

Canada

100-03945/

Canadian Technical Report of
Hydrography and Ocean Sciences 205

2000

Ocean Sciences Data Rescue Inventory

by

Anthony W. Isenor and Fasil Woldegeorgis[1]

Ocean Sciences Division
Martimes Region
Fisheries and Oceans Canada
Bedford Institute of Oceanography
P.O. Box 1006
Dartmouth Nova Scotia,
Canada
B2Y 4A2

[1]Evans Computer Applications Ltd.
6424 Norwood Street
Halifax, Nova Scotia
Canada
B3H 2L3

© Public Works and Government Services 1999
Cat. No. Fs 97-18/205 E ISSN: 0711-6764

Correct citation for this publication:

Isenor, Anthony W. and Fasil Woldegeorgis. 2000. Ocean Sciences Data Rescue Inventory, Can. Tech. Rep. Hydrog. Ocean Sci. 205, iv + 30p.

Table of Contents

Abstract iv

Résumé iv

1.0 Introduction 1

2.0 Metadata 2

2.1 Ship Schedules 2
2.2 Existing Databases 2
2.3 Previous Efforts at Identifying Missing Cruise Datasets 2
2.4 Cruise Reports and Sailing Instructions 3

3.0 Resulting Database - BCD 3

3.1 Known BCD Problems 7
3.1.1 Cruise Numbers 7
3.1.2 Marsden Squares 7
3.1.3 Data Types 7
3.1.4 Current Meter Data and Cruise Numbers 8
3.1.5 Cruise Report Source vs. Data Location 8
3.1.6 Missing Information in AllCruises 8

4.0 Current Database Characteristics 9

4.0 Summary of Existing Databases 9

4.1 AFAP (Climate) 9
4.2 Nutrient 10
4.3 CCD 10
4.4 OSD Datashop Archive 10

5.0 The Scope of the Missing Data 11

5.1 Water Sample Data 11
5.2 Other Data Types 12

6.0 Conclusions 13

7.0 Acknowledgements 13

Appendix A: At Risk Datasets 14

Abstract

Isenor, Anthony W. and Fasil Woldegeorgis. 2000. Ocean Sciences Data Rescue Inventory, Can. Tech. Rep. Hydrog. Ocean Sci. 205, iv + 30p.

A database of historic cruise metadata from the Bedford Institute of Oceanography has been created. This database was built primarily from planned ship sailing schedules, with additional information from Sailing Instructions and Cruise Reports. Cruise metadata were digitised, including scientist, start and end dates, general area of operation, and platform. Archived records were searched to identify types of data collected and the Marsden squares for each cruise. An attempt was made at comparing the data in existing databases, such as the Climate Database and various Chemistry databases, to the collected data types identified from the reports. A view of the present database state is provided in the following report.

Résumé

Isenor, Anthony W. and Fasil Woldegeorgis. 2000. Ocean Sciences Data Rescue Inventory, Can. Tech. Rep. Hydrog. Ocean Sci. 205, iv + 30p.

Le présent rapport fait état de la création d'une base de metadonnées historiques sur les croisières en partance de l'Institut océanographique de Bedford. Cette base de données a été établie principalement à partir des calendriers des croisières prévues, avec des renseignements additionnels tirés des instructions de navigation et des rapports de croisière. Les métadonnées sur les croisières ont été numérisées, y compris les noms des scientifiques, les dates du début et de la fin des croisières, les secteurs généraux des opérations et les plates-formes. On a fait des recherches dans les dossiers archivés pour identifier les types de données recueillies et les cartes de Marsden pour chaque croisière. On a essayé de comparer les données dans des bases de données existantes, dont la base des données climatiques et diverses bases de données à caractère chimique, aux types de données recueillies identifiées dans les rapports. Le présent rapport brosse aussi un tableau de l'état de la base de données actuelle.

1.0 Introduction

The oceanographic community's interest in large-scale problems has placed renewed attention on existing data. In particular, the research community has been working on the assimilation of datasets into existing databases. Such efforts tend to maximise the worth of the database systems, as well as the datasets. By using these historic datasets, researchers can address issues related to broader spatial and temporal scales.

In Ocean Sciences Division (OSD) and Marine Environmental Science Division (MESD) at the Bedford Institute of Oceanography (BIO), there has been considerable effort directed toward existing datasets. This interest and the resulting products have evolved over the past years to become an integral part of the science community's knowledge base. This evolution had its beginning in the creation of single user databases dealing with data of specific interest to the researcher. These databases quickly became of use to others in the oceanographic community, both locally and regionally. As this evolution continued, the databases became larger and recognised nationally.

More recently, these databases are being consolidated to address cross-discipline issues. As well, the access to the databases is changing as computer networks become more widely used. Access to the actual data, as opposed to products, is being developed providing the global community with the ability to analyse these historic data.

An obvious question in this evolution of databases relates to the volume of available historic data. The data management community must be able to identify those datasets not yet in the databases. Thus, a management tool was required to track the amount of collected data and ultimately compare this to the database archives. In terms of BIO, a list of BIO data collection operations from 1962 onward (the opening of the Institute) was required to provide the community with a means to compare collection operations with archived datasets.

The task of producing a database of BIO data collection operations was also initiated due to the paucity of water sample data in the Northwest Atlantic as compared to known data collection operations. The terse examination of a few missing cruise datasets identified the need for a more complete examination of collected datasets. The project objective was to assess how many datasets were in fact missing from the local databases. This process would identify data in need of rescuing and could then be used to direct subsequent data rescue efforts.

2.0 Metadata

Metadata is collectively referred to as "data about data". In this context, metadata attributes identify cruise specific information. For example, the platform of operation, the scientist in charge, start and end dates, the area of operation, etc.

The metadata associated with a cruise dataset is often in many locations. As well, a cruise dataset may be defined in many different ways. Using cruise numbers to define a dataset results in various problems, as discussed below.

The following sections list the metadata information sources for this project.

2.1 Ship Schedules

A near complete listing of BIO cruises existed in the Geoscience Centre at BIO (NRCAN). The paper listing of planned ship schedules covered the period 1961 to 1992. This listing included scientist, platform, start and end dates, and a description of the general area of operation.

For years after 1992, similar information was obtained from the charted ship schedules produced by the, now non-existent, BIO Ships Division.

2.2 Existing Databases

Several databases in use within OSD and MESD were also considered as information sources for cruise metadata. In particular, the chemistry cruise, the chemistry nutrient, and the AFAP databases, and the Datashop archive existing on the mass storage system were considered. These sources identified numerous cruise datasets that were not identified in the ship schedule list.

2.3 Previous Efforts at Identifying Missing Cruise Datasets

The chemistry group undertook the identification of cruises that were thought to collect nutrient data. By searching the cruise reports available in the BIO library, a list of possible cruises were identified. This list was cross-referenced with the created database to identify missing cruise numbers.

2.4 Cruise Reports and Sailing Instructions

Cruise reports pertaining to physical oceanographic cruises from within the Coastal and Ocean Circulation groups of OSD (and previous incarnations of this Division) are available from the OSD Datashop. These reports were examined for cruise number information, as well as used to compile a listing of collected data types for the particular cruise.

Other cruise reports and Sailing Instructions are available in the BIO library. These documents were also used to identify missing information.

3.0 Resulting Database - BCD

The creation of a digital copy of the ships schedule information was the starting point of the database. For lack of better terminology, and since this database focuses on a listing of BIO cruises, we refer to the created database as the BIO Cruise Database (BCD).

All cruises from the ship schedules were then cross-referenced with the other sources of information. This task helped ensure that the BCD truly contained a listing of all BIO cruises. This exercise identified numerous inconsistencies in the data sources. These inconsistencies typically dealt with the named chief scientist, start and end dates, and planned verses actual cruises. We attempted to rectify these inconsistencies by establishing a hierarchy of importance to the data source. Cruise reports were considered over Sailing Instructions, which were in turn considered over the ship schedules. Of course, the combinations of potential inconsistencies are so great, that in some cases educated reasoning was applied to the attribute in question.

After identifying all BIO cruises, the physical oceanographic cruise reports available from the OSD Datashop were reviewed. Information pertaining to the type of data collected and the location of the data were also including in the BCD.

The final BCD table structure consists of five tables outlined in Table 1. The field structure within these tables is outlined in Table 2.

Table Name	Description
AllCruises	All cruises from BIO are listed. This is based on ship schedule information, various databases, sailing instructions and cruise reports.
Cruise Report Source	Identifies the source of information used to identify the data types sampled.
Data Type	A look-up table of code names and long descriptions, identifying types of data collected. This list is very wide in scope, including instrument tests, etc. as data types.
Data Specs	Lists the data types collected on the particular cruise.
Marsden Square	Lists the location of the data collection for the cruise.

Table 1. Table names and a brief description.

Table Name	Field Name
AllCruises	**cruise_number** – the BIO assigned cruise number **leg** – subcruise designation **platform** – the platform used during data collection **start_date** – start date of the cruise **end_date** – end data of the cruise **location** – the area of focused data collection **chief_scientist** – chief scientist on the cruise **organizations** – section, division, etc. **notes** – general comments
Cruise Report Source	**cruise_number** – as above **leg** – as above **source of information** – the source of data type information
Data Type	**short_description** – short code for data type **long_description** – long description of the data type
Data Specs	**cruise_number** – as above **leg** – as above **data_type** – short description from "Data Type" **data_location** – where the data currently reside **data_state** – media on which data currently reside **rescue_effort** - effort code to get the data into an existing database
Marsden Square	**cruise_number** – as above **leg** – as above **Marsden_square** - a Marsden square where data was collected

Table 2. Table and field names. Field names have a brief description.

A schematic of the data flow within the tables is shown in Figure 1. The table Cruise Report Source uses AllCruises as a source for cruise numbers. In turn, Data Specs is a subset of the cruises listed in Cruise Report Source. Data Type is also used in Data Specs to identify the collected data for a particular cruise. The Marsden Square information is not related to data types, but rather cruises. Thus, Marsden Squares is related to Cruise Report Source, not Data Specs.

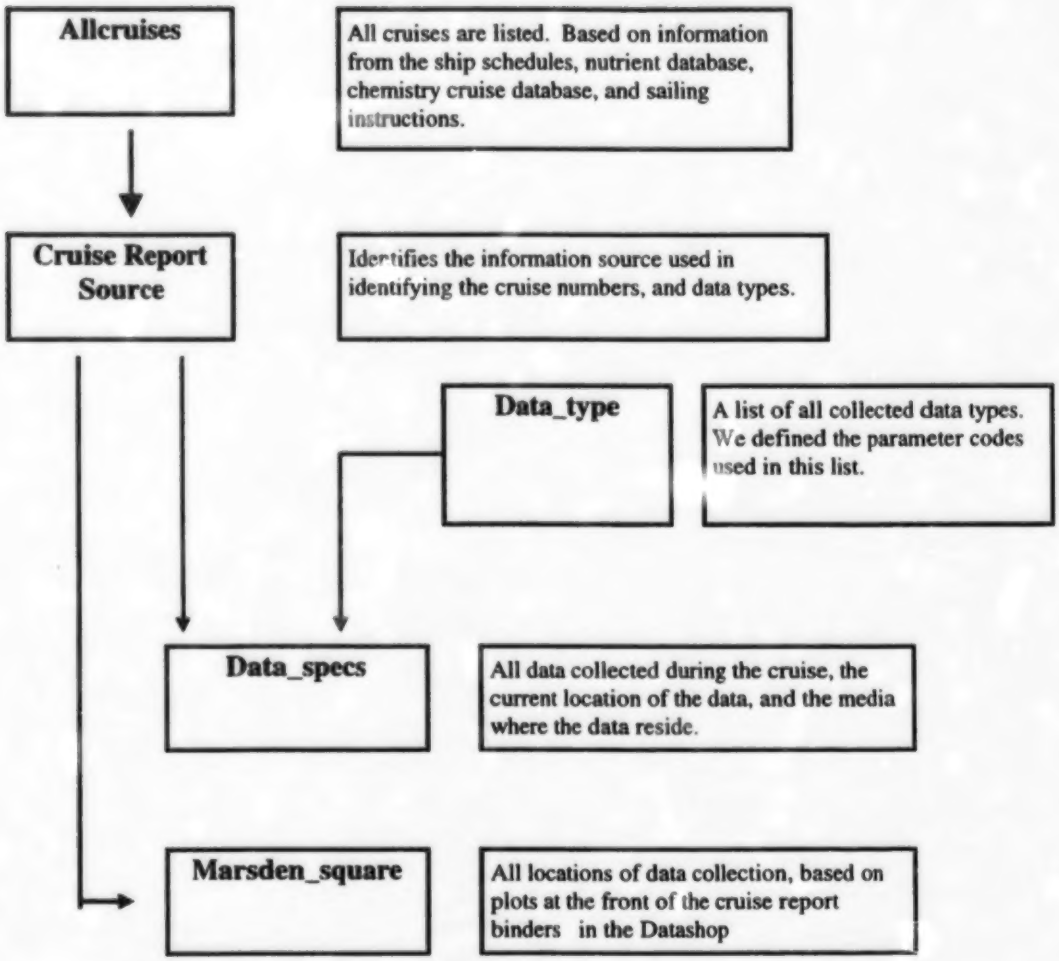

Allcruises
All cruises are listed. Based on information from the ship schedules, nutrient database, chemistry cruise database, and sailing instructions.
Cruise Report Source
Identifies the information source used in identifying the cruise numbers, and data types.
Data_type
A list of all collected data types. We defined the parameter codes used in this list.
Data_specs
All data collected during the cruise, the current location of the data, and the media where the data reside.
Marsden_square
All locations of data collection, based on plots at the front of the cruise report binders in the Datashop

Figure 1. Schematic of the data table structure.

3.1 Known BCD Problems

There are numerous problems with the information contained within these tables. These problems are discussed below.

3.1.1 Cruise Numbers

The technique of numbering a dataset based on cruise number presents some unique problems. Multiple cruise numbers occur frequently in the 1960's because of numerous groups assigning different numbers. For example, the Canadian Oceanographic Data Centre (a precursor to the Marine Environmental Data Services, MEDS) assigned a CODC report number, the group scheduling ship time assigned a ship related cruise number and BIO assigned another cruise number.

Another problem occurs when cruise numbers are defined internally by data administrators. This situation arises when datasets collected by other organisations are added to internal databases. Such cruise numbers, although not important for data rescue within OSD, were included in the created database. This was necessary to complete the listing of available datasets within the OSD databases.

3.1.2 Marsden Squares

The information in Marsden Square was based on numerous sources. When possible, the plots at the beginning of cruise report sections bound in the OSD Datashop were used. It is difficult to know what data types were collected in which Squares. This is compounded when along track data sampling is conducted, resulting in the ship track defining the Marsden Squares of data collection. In cases where there were no maps, longitude and latitude measurements from the cruise report or sailing instructions were used. In the case when only an area of interest was defined, the resulting definition of Marsden Squares may be an exaggeration of the actual sampling area.

3.1.3 Data Types

The Data Type table contains a wide scope of "types". This list includes instrument tests, hardware tests, and broad categories of sampling, such as "Geological samples".

A question mark field, "?" was added to indicate an unknown data type. This is used when we are aware of stored data, but do not know the details of the stored parameters.

3.1.4 Current Meter Data and Cruise Numbers

The number scheme used for current meter datasets is based on the deployment cruise number. When reviewing cruise reports, the distinction between mooring deployment and recovery is often unclear, and thus we expect some current meter data statistics to erroneous.

3.1.5 Cruise Report Source vs. Data Location

The Cruise Report Source table identifies the source of information for the data types specified in Data Specs. Of the information sources listed in the Cruise Report Source table, 88 sources are identified as CCD (Chemistry Cruise Database) and 17 sources based on the Strain/Bussard library search. From the 88 CCD sources, data types were identified for 46 cruises. For the 17 library search cruises, 12 have identified data types.

The Cruise Report Source lists the source of information for the data types. This is not to be confused with the field "data location" contained in the table Data Specs. The "data location" addresses the question "where would I go if I wanted to get that data?" The Cruise Information Source identifies how we initially became aware of the cruise and data types.

3.1.6 Missing Information in AllCruises

The list of cruise attributes in AllCruises is as complete as the information sources would allow. Priority was placed on identifying the cruise dates, location and platform. However, after all checks, some fields remain empty. Considering only those cruise numbers defined by actual BIO ship data collection operations (as opposed to including numbers based on external datasets not involving a BIO ship), the missing fields in AllCruises can be summarised as follows:

- 51 cruises do not have an identified starting date,
- 75 do not have an identified end date,
- 62 cruises do not have an identified senior scientist .
- 100 cruises do not have an identified location (area)
- 7 cruises do not have an identified platform.

Numerous cruise numbers listed in the ship schedules were identified as cancelled or with no additional information. These cruises were not included in the BCD.

4.0 Current Database Characteristics

The current characteristics of the database tables are as follows:

The AllCruises table contains 2351 records. These are all the unique BIO ship based data collection operations (totalling 1905) since the beginning of the Institute as well as other cruise numbers assigned by individual data administrators of non-ship operations (totalling 446). These 446 cruise numbers were included in the AllCruises table but appear with no other attributes.

There are 2351 records in Cruise Information Source. Of these, 992 records have an identified information source.

There are 3186 records in Data Specs, with 972 records being unique cruise numbers.

There are 154 data types.

There are 629 identified Marsden Squares, covering 293 unique cruises.

4.0 Summary of Existing Databases

The existing databases within OSD are expected to be the target locations for rescued datasets. Thus, it is worthwhile to understand what exists within these databases.

Note that these databases are used in the Data Location field of the Data Specs table. This identifies when the particular data type from a specific cruise is included in the database. By identifying what is included in the databases, we by default identify the missing data.

4.1 AFAP (Climate)

The AFAP climate database project is an ongoing effort to assemble all of the temperature and salinity data for the Northwest Atlantic.

The AFAP database stores temperature and salinity data from the following data sources: CTD, Batfish, water sample, various BT types, Tesac and Bathy messages. This list also indicates the data type hierarchy used to omit data that overlap in space and time. CTD data is given the highest level, and replaces all other data from the same space/time location.

4.2 Nutrient

The Marine Environmental Science Division maintains the nutrient database (referred to as the chemistry nutrient database). This database contains spatially and temporally referenced nutrient, oxygen, salinity, temperature and related data based on water sample measurements. The geographical focus area for the nutrient database is the Northwest Atlantic and Arctic. The database contains data from a large number of countries and international oceanographic data centres.

4.3 CCD

The CCD represents a diverse set of data types stored in a database maintained by the Marine Environmental Science Division. The stored data represent spatially and temporally referenced water sample and sediment parameters. This database contains a more varied group of data types, as compared to the nutrient database. The geographical focus area for the CCD is the Northwest Atlantic and Arctic. Most of the data in this database was originally produced at BIO.

4.4 OSD Datashop Archive

The OSD Datashop Archive refers to the data stored on the mass storage system (msm). This location, including both the temporary and permanent archives, contains the following data types:

- Batfish (BATF),
- Water sample data (BOTL),
- CTD (CTD),
- Drifting buoy (DRIB),
- Moored argos buoy (MAB),
- Moored current measurement (MCM),
- Moored CTD (MCTD),
- Meteorological (MET),
- Moored meteorological buoy (MMB),
- Moored temperature chain (MTC),
- Tide gauge (MTG),
- Moored thermograph (MTR),
- Profiling current meter (PCM),
- Temperature depth recorders (TDR) and
- XBT (XBT).

These data are organised by data type and year.

5.0 The Scope of the Missing Data

After completing the identification of data location, the data missing from the recognised databases becomes apparent. The following demonstrates some potential uses of the BCD.

5.1 Water Sample Data

As an example of the information within the BCD, consider the three water sample data parameters: silicate (used as an indicator of nutrient collection), oxygen and salinity. For this summary, the historical record was broken into pentads starting with the period January 1, 1960 to December 31, 1964. For each pentad, the percentages of the three parameters, based on a cruise dataset, in any of the recognised databases were determined. Figure 1 shows the graphical representation of the amount of data stored in the recognised databases for each pentad.

Several characteristics of the curves are apparent. The lack of data from 1960 to 1969 covers 10 unique cruises. Four of these cruises collected silicate data, eight collected oxygen data and nine collected salinity data.

A second feature is the drop in stored data in the mid-80s. This is thought to be a result of conversion in the computer operating systems, and in particular the transition involving NOS VE.

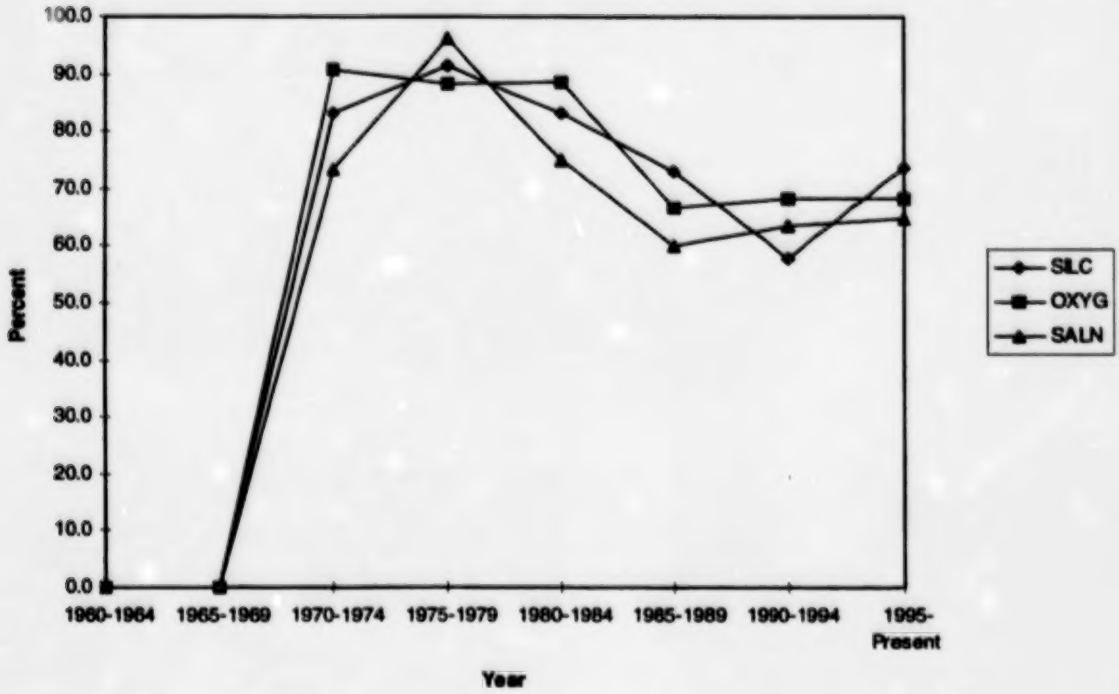


Figure 2. Plot showing the percentage of the indicated data type that is contained in one or more of the databases.

Appendix A uses these same three parameters to identify the missing data type by pentad and Marsden Square.

5.2 Other Data Types

The data specs table can be used to identify CTD data missing from the recognised databases. In total, 24 CTD datasets have not been located in the recognised databases. Of these 24, 19 are from 1993 or before. Later datasets may be in the process of being placed in a database. A search of the AFAP database indicated that these 24 datasets were not present.

Table 3 indicates several other data types and the number of missing datasets for that type.

Data Type	Number Missing	Number Missing from 1960-1993
CTD	24	19
ADCP	43	35
Batfish	38	35
Moored Current Meter	132	123

Table 3. Number of missing datasets for other data types.

6.0 Conclusions

This effort has produced a database that can be used to track datasets from collection to an existing database. We have identified numerous datasets that are at present not contained in existing databases. The next step of this project should be to prioritise and attempt to rescue these datasets.

We would propose that priority be placed on the water sample oxygen, nutrients, salinity and CTD data. Work would be conducted backward in time, from the most recent to the oldest dataset. The investigation would focus on the existence and state of the data. After determining the state, we would then rescue the data if possible. If rescue were not possible, we would document the attempted avenues and all findings. This may assist other rescue efforts.

For updating the BCD, we would propose this be the responsibility of the OSD Datashop. This group is currently responsible for OSD data processing and archival, and thus may use the BCD as a tool for tracking expected datasets. The BCD could also be extended to track processed datasets.

7.0 Acknowledgements

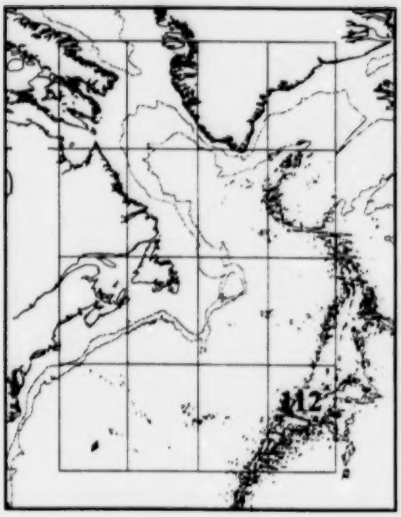
The Marine Environmental Data Service, using funds from Ocean Climate Program under the Department of Fisheries and Oceans supported this work.

Appendix A: At Risk Datasets

The following tables and maps identify some of the cruise datasets that have not been located in any of the recognised databases. The tables indicate the pentad, parameter and cruise number for a particular Marsden Square. The map indicates the location of the Marsden Square.

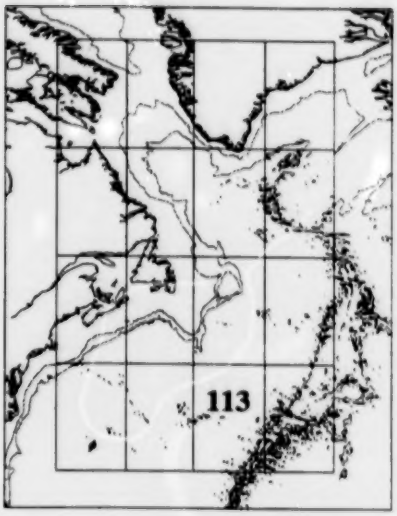
Note that the reason the dataset is missing has not been identified. Recent datasets may be missing from the databases because the processing has not been completed. Datasets since 1995 may be in this category.

Mardsen-112



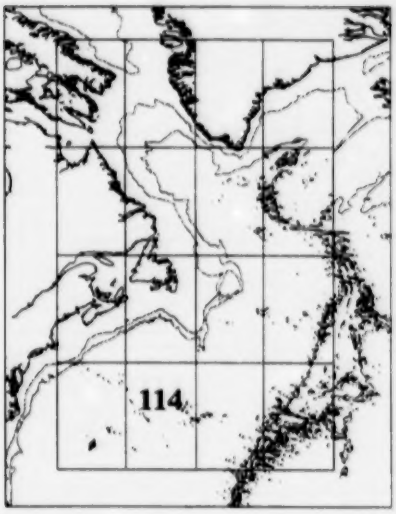
	Oxygen	Salinity	Silicate
1960-1964	0463 1263 0964	1263 0964	0964
1965-1969			
1970-1974			
1975-1979			
1980-1984			
1985-1989			
1990-1994	91007	91007	91007
1995-present			

Mardsen-113



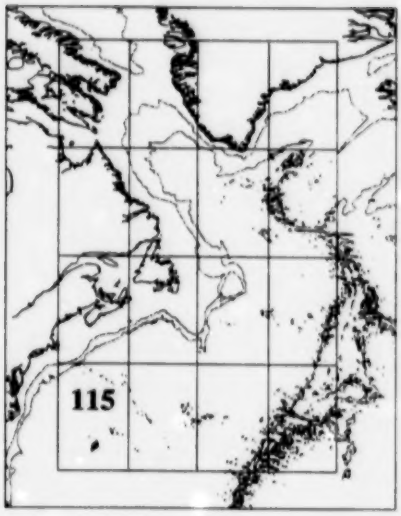
	Oxygen	Salinity	Silicate
1960-1964	0463 1263 0964	1263 0964	0964
1965-1969			69025
1970-1974	72008	71029 72008	70023 72008
1975-1979			
1980-1984	80029	80029	
1985-1989			
1990-1994	91007	91007	91007
1995-present			

Mardsen-114



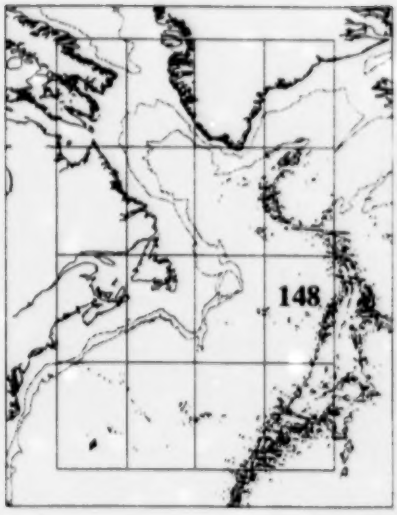
	Oxygen	Salinity	Silicate
1960-1964	0463 1263	1263	
1965-1969			
1970-1974			
1975-1979			
1980-1984	83029 84012	83029 84012	
1985-1989	88001		88001
1990-1994			
1995-present			

Mardsen-115



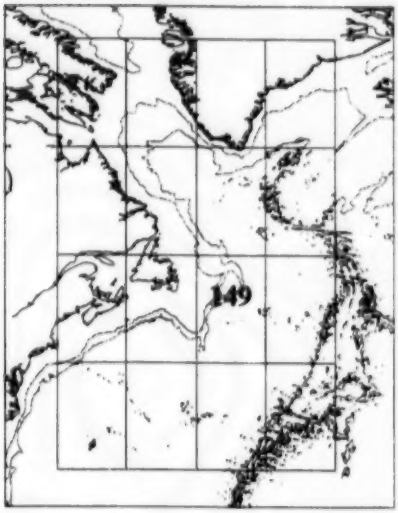
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969			
1970-1974			
1975-1979			
1980-1984	83029 84012	83029 84012	
1985-1989			
1990-1994			
1995-present			

Mardsen-148



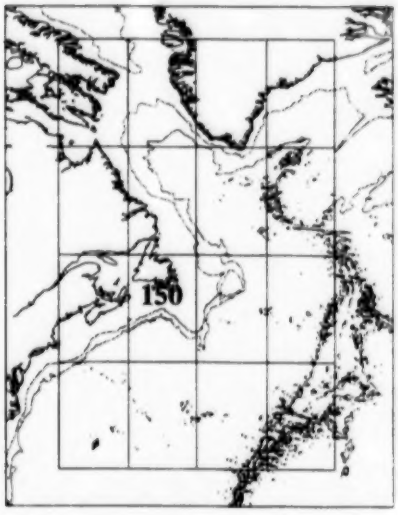
	Oxygen	Salinity	Silicate
1960-1964	0964	0964	0964
1965-1969	0566	0566	69025
1970-1974			
1975-1979			
1980-1984			
1985-1989			87037
1990-1994	91007	91007	91007
1995-present			

Mardsen-149



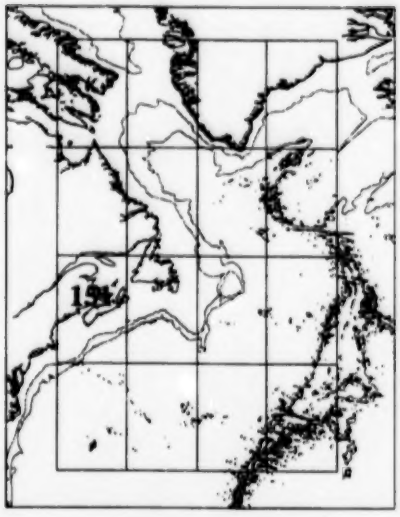
	Oxygen	Salinity	Silicate
1960-1964	0463 1263 0964	1263 0964	0964
1965-1969	266 566	266 566	69025
1970-1974	72008	71029 72008	70023 72008
1975-1979		79001	
1980-1984		80002 81024	81007 81024
1985-1989	88025 89037	86031 87012 89005 89037	87037 89037
1990-1994	91007	91007 92023	91007
1995-present			

Mardsen-150



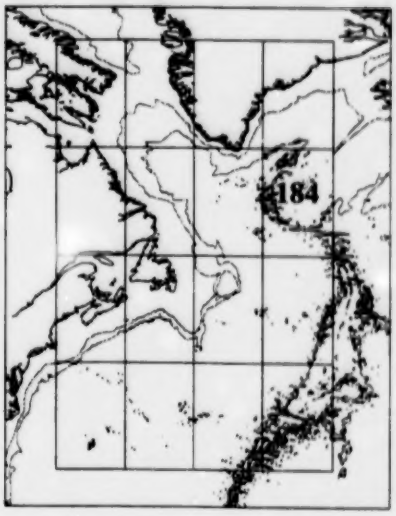
	Oxygen	Salinity	Silicate
1960-1964	463 S64	S64	
1965-1969		2968 4868	
1970-1974		74020	
1975-1979			
1980-1984	83029 84012	83029 84012	84048
1985-1989	88001 88025	85032 85040 86031 87012 89041	85017 85039 86037 88001 89036
1990-1994	91007	91007	90030 91007 91058 93040 94033
1995-present	96006, 96014 96026 97009 98023	96006, 96014 96026 97009 98023	96006, 96014 96026 97009 98023

Mardsen-151



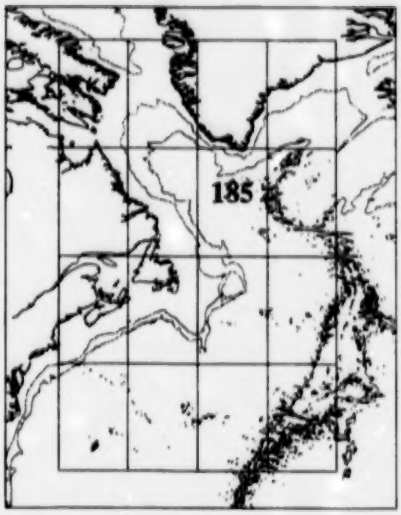
	Oxygen	Salinity	Silicate
1960-1964	S64	S64	
1965-1969	566	566 2968 4868	
1970-1974		74001	
1975-1979	78001		78001
1980-1984	83029 84012	81042 82007 83029 84012	83038 84048
1985-1989	87029 88041	85032, 85040 86025, 87029 88041	85017, 85039 85041, 86037 89036
1990-1994	90025 91007 94018	91007 91028 94018 94041	90025, 90030 91007, 91058 92050, 93033 93040, 94029 94033, 94041
1995-present	95034, 96006 96014, 96026 97009 98023	95010, 95034 96006, 96014 96026, 97009 98023	96006, 96014 96026 97009 98023

Mardsen-184



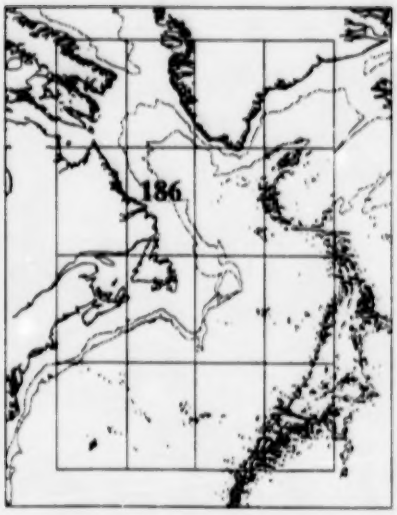
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	0266	0266	
1970-1974			
1975-1979			
1980-1984			
1985-1989			
1990-1994	91007	91007 93027	91007
1995-present			

Mardsen-185



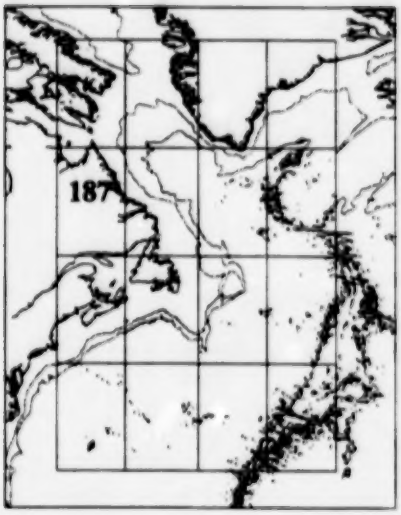
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	0266	0266	
1970-1974			
1975-1979			
1980-1984			
1985-1989	86021 88025	86021	86021
1990-1994	91007	91007 92023 93027	91007
1995-present	96026 97009	96026 97009	96026 97009

Mardsen-186



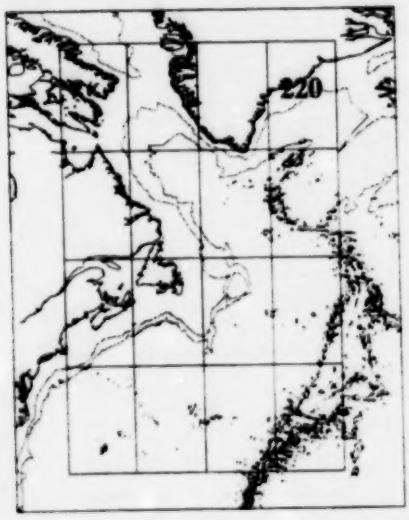
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	0266	0266	
1970-1974			
1975-1979	77029		77029
1980-1984		80026	80026
1985-1989	86021 87026 88025	86021 87026	86021 87026 89036
1990-1994	91007	91007	91007
1995-present	96006 96026 97009 98023	96006 96026 97009 98023	96006 96026 97009 98023

Mardsen-187



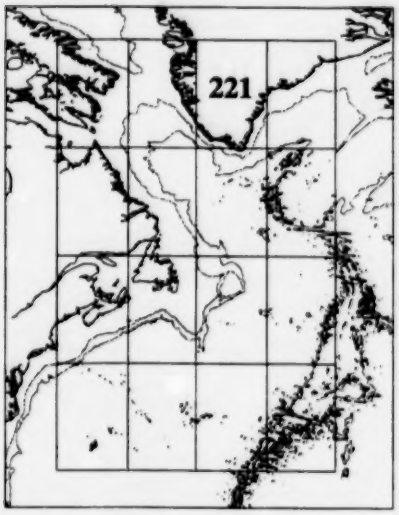
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969			
1970-1974			
1975-1979	77029		77029
1980-1984			
1985-1989		86029	
1990-1994			
1995-present			

Mardsen-220



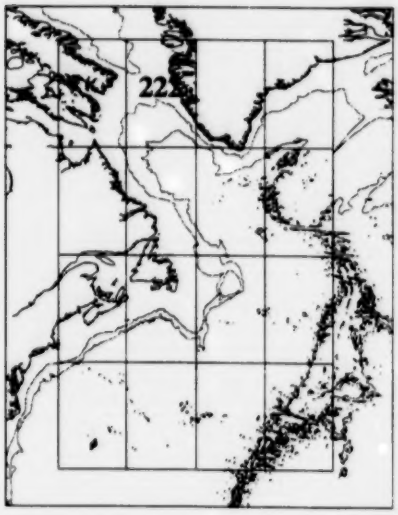
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	267	267	267
1970-1974			
1975-1979			
1980-1984			
1985-1989			
1990-1994		93027	
1995-present			

Mardsen-221



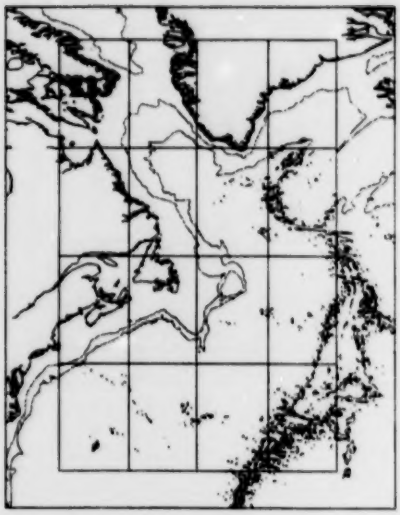
	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	266	266	
1970-1974			
1975-1979			
1980-1984			
1985-1989			
1990-1994			
1995-present	96006 96026 97009 98023	96006 96026 97009 98023	96006 96026 97009 98023

Mardsen-222



	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	2265 266	2265 266	2265
1970-1974			
1975-1979			
1980-1984			
1985-1989	88037 89016	87031 88037	89016
1990-1994	90022	90022	90022
1995-present	96026 97009	96026 97009	96026 97009

Mardsen-223



	Oxygen	Salinity	Silicate
1960-1964			
1965-1969	2265	2265	2265
1970-1974			
1975-1979			
1980-1984			
1985-1989	88037	86029 87031 88037	
1990-1994	90022	90022	90022
1995-present			